Unbiased statistical testing of shared genetic control for potentially related traits

Chris Wallace

Department of Medical Genetics

JDRF/Wellcome Trust Diabetes and Inflammation Laboratory

NIHR Cambridge Biomedical Research Centre

Cambridge Institute for Medical Research

University of Cambridge

Wellcome Trust/MRC Building

Cambridge

CB2 0XY

UK

chris.wallace@cimr.cam.ac.uk

Abstract

Integration of data from genomewide single nucleotide polymorphism (SNP) association studies of different traits should allow researchers to disentangle the genetics of potentially related traits within individually associated regions. Methods have ranged from visual comparison of association p values for each trait to formal statistical colocalisation testing of individual regions, which requires selection of a set of SNPs summarizing the association in a region. We show that the SNP selection method greatly affects type 1 error rates, with all published studies to date having used SNP selection methods that result in substantially biased inference. The primary reasons are twofold: random variation in the prescence of linkage disequilibrium means selected SNPs do not fully capture the association signal, and selecting SNPs on the basis of significance leads to biased effect size estimates.

We show that unbiased inference can be made either by avoiding variable selection and instead testing the most informative principal components or by integrating over variable selection using Bayesian model averaging. Application to data from Graves' disease and Hashimoto's thyroiditis reveals a common genetic signature across seven regions shared between the diseases, and indicates that for five out of six regions which have been significantly associated with one disease and not the other, the lack of evidence in one disease represents genuine absence of association rather than lack of power. Our detailed examination of the statistical properties of colocalisation tests and associated software will foster more widespread adoption of formal colocalisation testing, especially given the increasing availability of large expression and genetic association datasets from disease-relevant tissue and puried cell populations which, coupled with identification of regulatory sequences by projects such as ENCODE, has the potential to reveal both shared genetic signatures of related traits and causal disease genes and tissues.

Introduction

In recent years, genomewide association studies (GWAS) have facilitated a dramatic increase in the number of genetic variants associated with human disease and other traits such as gene expression. Understanding the means by which these variants exert their effect will aid the design of the detailed functional followup studies already underway. Although the causal variants are not commonly known, multiple traits have been mapped to the same genetic loci, raising the possibility that

the same variants affect multiple traits either directly or with one trait mediating the other. For example, genetic susceptibility to type 2 diabetes across 12 loci appears mediated by the genetic influence on body mass index [1]. Within individual loci, researchers are examining the genetic association signals from pairs of traits in parallel, with similar results interpreted as evidence that the two traits may colocalise, or share a common causal variant. These traits may be eQTL signals across two or more tissues [2, 3], eQTL and disease signals [4, 5] or two or more diseases [6]. Distinguishing cases where related diseases share a common causal variant *versus* those where neighbouring but distinct variants appear to underly disease risk in a region will aid identification of cross-disease and disease-specific mechanisms. In addition, comparison of disease and eQTL data has the potential to reveal both the likely disease causal gene in regions where extensive linkage disequilibrium (LD) and/or gene density means a number of candidate causal genes exist, and the relevant tissue type where tissue-specific eQTLs exist. However, dependence between genotypes at neighbouring SNPs, caused by LD, means that determination of colocalisation is not obvious, as there may exist distinct but neighbouring causal variants for each trait which are mutually associated.

When these traits are measured in the same individuals, standard statistical techniques, for example, conditioning on one trait, may be used to determine whether one trait mediates the other [1]. However, when they are measured in distinct samples, most researchers have approached the task of looking for colocalisation either by examining by eye the association signals across a set of common SNPs in the two datasets [7] or by testing for evidence of residual association in their available dataset conditional on the most associated SNP in the other [4]. When full data for both traits are available, colocalisation may be tested directly by examining whether coefficients from regressions of each trait against two or more SNPs are proportional, as they should be if those SNPs jointly tag a common causal variant [5, 8].

We show here that naive application of both conditional and direct colocalisation tests may result in substantially inflated type 1 errors, and explore reasons for that inflation. The inflation cannot be easily resolved for conditional tests, but we demonstrate two alternative approaches for direct testing which result in unbiased inference. Finally, we apply these methods to colocalisation

testing of 13 regions shown in Supplementary Table 1 which have been associated with one or both of the autoimmune thyroid diseases, Graves' disease (GD) and Hashimoto's thyroiditis (HT), using previously published dense genotyping data [9].

Methods

Statistical colocalisation testing

We begin by introducing some notation. Assume two traits, Y and Y', have been measured in distinct samples and evidence for association of both traits has been identified to some genetic region. Let the region be covered by p SNPs genotyped in both samples, with the genotype matrices denoted by $X = (X_1, \ldots, X_p)$ and $X' = \left(X'_1, \ldots, X'_p\right)$ respectively. Conditional approaches begin with identifying the most strongly associated SNPs for Y and Y', SNPs k and k', say, then examine whether there is any evidence for association between Y and SNP k conditional on SNP k'. The null hypothesis is therefore

$$H_0^{\text{cond}}: \quad Y \perp\!\!\!\perp X_k | X_{k'} \tag{1}$$

Concerned that LD would make interpration of the conditional test difficult, Nica et. al [4] extended the conditional method as follows. For every SNP j generate residuals Y_j^R from a regression of Y against X_j and test the correlation of Y_j^R and X_k using Spearman's rank correlation test, generating p values P_j. The evidence against the null hypothesis (1) is then measured by the rank of $P_{k'}$ in the empirical distribition, $[P_j]$, generated. This effectively compares the p value at the test SNP k conditional on SNP k' to that conditioning on all other SNPs in the region.

The direct approach frames the null hypothesis differently. A set of q SNPs are chosen which are deemed somehow to jointly be good predictors of one or both traits. Regressing Y and Y' against these columns of X and X' respectively produces estimates, $\mathbf{b}_1$ and $\mathbf{b}_2$, of regression coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, with variance-covariance matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ respectively. Since sample sizes are large, the combined likelihood may be closely approximated by a Gaussian likelihood for $(\mathbf{b}_1, \mathbf{b}_2)$, assuming $\mathbf{V}_1, \mathbf{V}_2$ are known and that $\text{Cov}(\mathbf{b}_1, \mathbf{b}_2) = 0$. Assuming equal LD in the two cohorts, ie that the

correlation structure between the SNPs does not differ, the null hypothesis is

$$H_0^{\text{prop}} : \boldsymbol{\beta}_1 \propto \boldsymbol{\beta}_2,$$

ie $\boldsymbol{\beta}_1 = \frac{1}{\eta}\boldsymbol{\beta}_2 = \boldsymbol{\beta}$ [8]. The chi-squared statistic

$$X(\eta)^2 = \mathbf{u}^T \mathbf{V}^{-1} \mathbf{u} \sim \chi^2 \tag{2}$$

is derived from Fieller's theorem [10], where $\mathbf{u} = \left(\mathbf{b}_1 - \frac{1}{\eta}\mathbf{b}_2\right)$ and $\mathbf{V} = \mathbf{V}_1 + \frac{1}{\eta^2}\mathbf{V}_2$. If η were known, $X(\eta)^2$ would have a χ^2 distribution on q degrees of freedom. Instead, Plagnol et. al take a profile likelihood approach and replace η by its maximum likelihood estimate, $\hat{\eta}$, which also minimises $X(\eta)^2$. Asymptotic likelihood theory suggests that $X(\hat{\eta})^2$ has a χ^2 distribution on $q-1$ degrees of freedom. Alternatively, one may take a Bayesian approach, and integrate the p value for $X(\eta)$ over the posterior distribution of η, generating posterior predictive p values [5]. The two are almost identical in large samples, but can differ in smaller samples, and posterior predictive p values have a somewhat different interpretation than standard p values [11, 12]. However, they avoid assuming the log-likelihood for η is approximately quadratic near its maximum which is not always the case; indeed it may be bimodal. The Bayesian approach assumes a Cauchy$(0, k)$ prior distribution for $\tan(\eta)$ and the effect of varying the parameter of the Cauchy distribution on the posterior predictive p values has been shown to be neglibable [5] and in this study we have set it to $k = 1$, which is equivalent to an uninformative prior on the $\tan(\eta)$ scale.

Simulation

We used simulation to demonstrate the effects of variable selection methods on the power and type 1 error rate for colocalisation testing. Full details are given in supplementary material. Briefly, we generated haplotypes from SNPs with a minor allele frequency of at least 5% found in phased 1000 Genomes Projcet data from the CEU population [13] across all 49 genomic regions outside the major histocompatability complex (MHC) which have been identified as type 1 diabetes (T1D) susceptibility loci to date [14]. These represent a range of region sizes and genomic topography

typical of GWAS hits. We excluded the MHC region which is known to have high variation, strong LD and exhibits huge genetic influence on autoimmune disease risk involving multiple loci and hence requires individual treatment in any GWAS [15]. Using a "causal variant" SNP chosen at random, we sampled case and control haplotypes according a multiplicative disease susceptibility model with relative risks of ranging from 1.1 to 1.3 to represent GWAS data or simulated a Gaussian distributed quantitative trait for which the causal variant SNP explains 30% of the variance to represent a moderately strong eQTL [3]. We then either used all SNPs or the subset of SNPs which appear on the Illumina HumanOmniExpress genotyping array to conduct colocalisation testing to reflect the scenarios of very dense targetted genotyping *versus* a less dense GWAS chip. All analyses were conducted in R [16] using the `coloc` package for direct colocalisation testing.

Colocalisation testing for autoimmune thyroid disease

An association study of the autoimmune thyroid diseases GD and HT has recently been completed using the Immunochip for genotyping, which provides dense coverage of regions of the genome previously associated with autoimmune disease [9]. The paper presented a total of 2285 Graves' disease cases, 462 Hashimoto's disease cases and 9364 controls. We split the controls randomly into two groups of size 4682, and analysed each of the 13 regions reported to be associated with one or both diseases [9]. We conducted direct colocalisation analysis of these coefficients using the the two alternative methods set out below.

Results

Naive application of colocalisation tests leads to biased inference

The choice of SNPs to use for testing can induce bias for two reasons. First, selecting the "most associated" SNP on the basis that the evidence for its association is strongest amongst all SNPs tested does not guarantee either that it is the causal SNP or even the best proxy. Random variation and LD mean that evidence for association may peak at an alternative SNP even when the causal SNP is included in the genotyping panel, a bias which is more pronounced for weaker effects and

smaller sample sizes (Supplementary Figure 5). Second, although it is well known that regression coefficients are unbiased estimates of population effects, this property does not hold after variable selection [17], an effect which has been referred to as "Winner's curse" in genetics [18, 19]. Choosing SNPs on the basis of their significance or some other measure of strength of association induces a bias away from the null - ie coefficients of selected SNPs are expected to overestimate the true effect - and again the effect is more pronounced for smaller sample sizes and weaker effects (Figure 1). These two biases mean that, in conditional testing, there is likely to be some residual association between the phenotype Y and the remaining genetic markers after conditioning on the selected SNP k' because (1) the conditioning SNP k' may not capture all the true association and (2) the estimated effect at the tested SNP k tends to be an overestimate. The result is very poor control of the type 1 error rate (Figure 2, track C1). Conditioning on the common causal variant rather than the most associated SNP (which is only possible in simulation studies) reduces the bias by removing the SNP selection problem, but does not eliminate it due to the overestimation of effect size (Figure 2, track C2).

As seen in Supplementary Figure 6, Nica's score tends towards to 1 for traits that share a causal variant and is uniformly distributed on $[0, 1]$ for distinct unlinked causal variants. Its distribution is increasingly skewed towards 1 as the LD between distinct causal variants increases. This makes sense if one considers that the case of two distinct variants in some LD lies partway between the extreme cases of distinct linked causal variants and a single common causal variant, which is equivalent to distinct causal variants in complete LD. The effect of using the most associated SNPs for testing compared with using the true causal SNPs is to reduce the skew towards higher rank scores as the r^2 between variants increases. Thus, while Nica *et. al*'s extension [4] is useful if searching for most likely colocalisation signals within a set, but as it avoids formally testing a null hypothesis, and because the scale against which to interpret the rank score is likely varies according to effect size, it does not provide a means to assess evidence for or against colocalisation at a given locus of interest.

For the direct approach, two strategies have been applied. Either colocalisation has been tested using the pair of SNPs k and k' defined above [8] or a lasso approach, where SNPs are first selected

in a lasso for one trait, and then additional SNPs are selected in a further lasso for the other trait [5]. This second approach was adopted in the hope it would reduce any bias. However, as shown by Miller [17], any variable selection method must induce bias in the estimated coefficients ($\mathbf{b}_1$, $\mathbf{b}_2$) if the estimation occurs in the same dataset as the selection, and we show here that neither method maintains control of the type 1 error rate (Figure 2, tracks D1, D2 and D3), although the bias is less extreme than for conditional testing. The lasso selection defined above does reduce bias compared to independent lasso selection in the two datasets, but, perhaps counter-intuitively, leads to greater bias than simply testing the pair of most associated SNPs (k, k') when only tagging genotypes are available and effect sizes are large (relative risk ~ 1.3). This is because, in this situation, lasso may select SNPs which are apparently weakly associated (either truly or through random noise) at which, as demonstrated in Figure 1, effect estimates are more strongly biased.

Proper control of type 1 error rates

The aim of selecting the most informative subsets of SNPs for direct colocalistion testing is to minimise the degrees of freedom of the test, and hence maximize power. However, unless independent data are available for variable selection, this increase in power comes at a cost to type 1 error rate control as shown above. We propose two methods for avoiding this problem.

If the region of interest displays strong LD, a modest number of principal components (PCs) are generally required to capture most of the SNP variation (Supplementary Figure 7) and we can perform the direct association test using a subset of the most informative components. Because PCs are by definition uncorrelated, and because the selection is not based on their relationship to the traits of interest, the estimated coefficients at any such subset are unbiased. Alternatively, we may combine the ideas of Bayesian model averaging (BMA) [20] and posterior predictive p values, to treat the model describing the joint association itself as a nuiscance parameter, and average the posterior predictive p values not just over the posterior for η, but also over the posterior for all SNP selection models. We found that, assuming a fixed or small number of SNPs (eg two or three) are sufficient to capture the association of both traits. Approximating the posterior probabilities by means of the Bayesian Information Criterion approximation [21, 22] and discarding highly improbable models

at the outset, this could be done without excessive computational burden (see Supplementary Material for full details). Both methods are available in our R [16] package, coloc, available from the Comprehensive R Archive Network (`http://cran.r-project.org/web/packages/coloc`).

Importantly, simulation shows either approach maintains good control of the type 1 error rate (Figure 3), even tending to be slightly conservative for small effect sizes or sample sizes. For PCs, as more components are selected, more information about the genetic variation in a region is captured, and hence we are more likely to accurately capture the signal of any causal variants. However, successive components add decreasing amounts of information whilst still adding another degree of freedom. At some point the negative effect of increasing degrees of freedom will outweigh the positive effect of increasing information, and we were concerned that the optimal test may depend heavily on the threshold used to determine the number of components selected. Instead, power seemed broadly acceptable once components capturing 70-90% of variation were selected (Supplementary Figure 8. In our 49 test regions, 70% of the variance could be captured by selecting an average of 7 (range 2-18) or 9 (range 3-44) components under a tagging or complete genotyping approach.

We compared power to detect departure from colocalisation using PCs capturing 90% of the genetic information, or the BMA approach to average over all two SNP models, or, to examine the theoretical maximum, using the causal variants themselves, which are known in a simulation study. When causal variants are known, power increases with sample size and effect size, but is negatively correlated with the r^2 between the causal variants, and is maximum when the two are completely unlinked ($r^2 = 0$). Both PC and BMA approaches have less power, reflecting the loss of information in not knowing these causal variants. Perhaps surprisingly given the differences in degrees of freedom for the two approaches, power was broadly similar, the BMA approach tended to perform slightly better.

Application to colocalisation testing of Autoimmune Thyroid Diseases

Existing evidence suggests that a single locus may contain variants which predispose to any one of multiple diseases, eg the non-synonymous C1858T SNP in *PTPN22* which is associated with

rheumatoid arthritis and T1D [23, 24], or distinct variants which predispose to different diseases, eg distinct variants in *IL2RA* are associated with T1D and multiple sclerosis [25, 26]. We used the direct colocalisation approach outlined above to examine the disease signals for the autoimmune thyroid diseases HT and GD from a recent dense genotyping study [9].

We first examined the seven regions where a significant single SNP effect has been identified in both diseases, ie at genomewide significant levels for GD and at a nominal significance threshold of $p < 0.05$ for HT, none of which display any evidence against colocalisation (Figure 4, all posterior predictive $p > 0.01$). The coefficient of proportionality, η, can be usefully interpreted when analysing two diseases. Two values of particular interest are $\eta = 0$ which would indicate no effect in HT given an effect in GD and $\eta = 1$ which indicates equal effects in each disease. In six of the seven regions, the credible interval for η includes 1, the exception being 2q33.2, where it is $(0.14, 0.99)$ or $(0.16, 1.03)$ under the PC and BMA approaches respectively. A value of $\eta \neq 1$ would indicate a stronger effect for one disease than another, but for 2q33.2 the interval is so close to 1, and the Bayes Factor leans towards favouring $\eta = 1$ (BF=5 or 7 under the PC and BMA approaches respectively) that this conclusion cannot be drawn with any confidence.

Turning to the the six regions where there is evidence of association in only GD, ie at genomewide significant levels for GD but $p > 0.05$ for HT in the single SNP analysis of these data [9]), we do not expect to see any departure from the null of colocalisation, without evidence of association to both traits. However, we can use η to infer whether this reflects a lack of power or genuine absence of association for HT. We evaluated the credible intervals for η in each region and compare the hypotheses of $\eta = 0$ and $\eta = 1$ using Bayes Factors. Across all six regions, the credible interval for η invariably includes 0 and the Bayes Factor either favours $\eta = 0$, or cannot distinsguish the hypotheses (Figure 4). Only for 6q27, which contains the candidate gene *CCR6* and has also been associated with Crohn's disease [27], rheumatoid arthitis [23] and vitiligo [28], is there any suggestion that $\eta \neq 0$ (the log Bayes Factor is marginally above 0 and $\hat{\eta}$ is closer to 1 than 0), which would be consistent with a shared association which has not yet been detected in HT.

Discussion

There are two sources to the bias in colocalisation testing presented above. The problem of variable selection is well studied in statistics generally [17] but has perhaps been neglected in statistical genetics, where the aim has been to detect convincing association to a region, rather than pinpoint the causal variant, particularly as most datasets to date have included an incomplete selection of variants in any region. Selecting SNPs which do not fully capture the trait association will affect conditional colocalistation testing because some residual association must remain after conditioning. On the other hand, it should not bias direct testing as the aim there is to test for proportionality of effect size rather than evidence of residual effect. This may explain the substantially higher error rates for naive conditional testing *versus* naive direct testing seen in Figure 2.

The bias in effect size estimates affects both methods, however. In genetics, we are familiar with "Winner's curse", which causes effect estimates which are examined conditional on the associated p value being below some significance threshold to be biased away from the null. Some attempts to correct this effect size bias have been made, either by modelling a selection procedure defined as a single SNP exceeding a predetermined level of significance [29], or by bootstrapping which can in theory account for the full selection strategy [30]. We explored both approaches, but found neither led to unbiased or even nearly unbiased inference (data not shown). Our proposed solution is to use direct testing and either avoid variable selection altogether by using the PCs which capture the majority of genetic variation in the region under test, or integrate over the variable selection using BMA. Either method maintains type 1 error, and the BMA approach appears to outperform the PC approach, although both have reduced power compared to the hypothetical scenario of being able to test the causal variants themselves.

As an example application of our proposed unbiased approaches to direct colocalisation testing, we analysed 13 loci associated with the autoimmune thyroid diseases GD and HT, and showed that in the seven regions where a locus has been associated with the two diseases, the data are consistent with common causal variants exerting an equal effect on each disease. In regions previously signficantly associated with only one disease, posterior predictive p values are unlikely to detect any departure from colocalisation but Bayes Factors comparing specific values of η can be

useful. Given the relatively smaller number of HT cases (462) compared to GD (2285), it might be expected that many of the loci only associated with GD have failed to be reach significance for HT due to lack of power. Estimates of power to detect association with HT under the assumption of equal effects in GD and HT are broadly similar across the seven regions associated with both diseases and the six regions associated with GD only (Supplementary Table 1), but these are likely to be over optimistic due to the expected bias in the GD effect size estimates. However, for five of these six regions, the evidence suggests that the association is genuinely specific to GD and further study of these loci in HT would be fruitless.

There is a pressing need for more widespread use of formal colocalisation testing. Researchers are turning to eQTL data to interpret GWAS results by simply considering whether an eQTL SNP is associated with any disease [31], visually [32], by conditional testing [4], by naive application of a direct colocalisation test [8, 5] or by attempting to integrate disease and gene expression association signals in networks [33]. Where colocalisation tests have been naively applied, we expect the null hypothesis of colocalisation has been rejected too readily, although this will affect loci with small and moderate effects to a greater degree than those with large effects. Thus, for our earlier analysis of colocalisation between T1D and monocyte gene expression signals [5], the list of loci compatible with colocalisation are likely correct, but some loci were probably erroneously rejected as potentially colocalising, and re-analysis of these data will be required.

In the case of network analysis, results can be difficult to reconcile with simple representations of the data. For example, lung expression and asthma genetic association data were integrated leading to the identification of *GSDMA* as the most likely causal gene for the asthma association in the 17q21 region [33], despite a graphical representation of the data showing that the SNPs most strongly associated with *GSDMA* expression were relatively weakly associated with asthma, and, *vice versa*, that the SNPs most strongly associated with asthma showed relatively weaker levels of association with *GSDMA* expression compared to the strongest signals. The asthma association on the 17q21 region was one of the first cases of explicitly using expression data to interpret disease association, with the association with asthma initially attributed to *ORMDL3* based on expression data from EBV transformed cell lines [34] and subsequently to *GSDMB* from a reanalysis of the

same data [35]. Candidate gene hypotheses have been constructed for all three genes. The lung expression data have greater potential for revealing the underlying gene, but, to hold confidence in results of analyses, particularly when the results contrast with simple visual inspection of the data, requires careful examination of the properties of the statistical method used.

Given the tissue-specific nature of many eQTLs identified to date [2, 3], there is a need for more large, publicly available eQTL datasets in a variety of disease relevant tissues and purified cell subsets to support the interpretation of existing GWAS data. Although expression data is typically shared after the publication of an eQTL study, we note that the genetic data must also be made available to allow full integration of eQTL and disease signals at shared loci. The increasing abundance of substantial GWAS datasets and the increasing availability of large eQTL datasets [3, 36, 33], together with our detailed examination of the statistical properties of colocalisation tests, the reassurance that it is possible to conduct these tests whilst maintaining type 1 error rates and the availability of software in our R package will facilitate more widespread formal colocalisation testing. Integration of genetic association data has the potential to refine understanding of underlying genetic mechanisms and aid in the design of follow-up studies already underway.

Acknowledgments

We thank Matthew Simmonds, Stephen Gough, Jayne Franklyn, Oliver Brand, for sharing their AITD genetic association dataset and all AITD patients and control subjects for participating in this study. The AITD UK national collection was funded by the Wellcome Trust.

Phased 1000 Genomes data was downloaded from http://www.sph.umich.edu/csg/abecasis/MACH/download/1

We would like to thank the UK Medical Research Council and Wellcome Trust for funding the collection of DNA for the British 1958 Birth Cohort (MRC grant G0000934, WT grant 068545/Z/02). We acknowledge use of data from The UK Blood Services collection of Common Controls (UKBS collection), funded by the Wellcome Trust grant 076113/C/04/Z, by the Wellcome Trust/JDRF grant 061858, and by the National Institute of Health Research of England. The DNA collection was established as part of the Wellcome Trust Case-Control Consortium.

For additional detailed acknowledgements of the source assay design, genotyping, data and

samples, please see Cooper et al (2012).

CW is funded by the Wellcome Trust (089989). The Diabetes and Inflammation Laboratory is funded by the JDRF, the Wellcome Trust and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. The Cambridge Institute for Medical Research (CIMR) is in receipt of a Wellcome Trust Strategic Award (100140).

References

[1] Li S, Zhao JH, Luan J, Langenberg C, Luben RN, et al. (2011) Genetic predisposition to obesity leads to increased risk of type 2 diabetes. Diabetologia 54: 776–782.

[2] Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325: 1246–1250.

[3] Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, et al. (2012) Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of hla alleles. Nat Genet 44: 502–510.

[4] Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, et al. (2010) Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations. PLoS Genet 6: e1000895.

[5] Wallace C, Rotival M, Cooper JD, Rice CM, Yang JHM, et al. (2012) Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. Hum Mol Genet 21: 2815–2824.

[6] Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, et al. (2011) Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet 7: e1002254.

[7] Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42: 295–302.

[8] Plagnol V, Smyth DJ, Todd JA, Clayton DG (2009) Statistical independence of the colocalized association signals for type 1 diabetes and rps26 gene expression on chromosome 12q13. Biostatistics 10: 327–334.

[9] Cooper JD, Simmonds MJ, Walker NM, Burren O, Brand OJ, et al. (2012) Seven newly identified loci for autoimmune thyroid disease. Hum Mol Genet 21: 5202–5208.

[10] Fieller EC (1954) Some problems in interval estimation. Journal of the Royal Statistical Society Series B (Methodological) 16: 175–185.

[11] Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Statist 12: 1151–1172.

[12] Meng XL (Sep., 1994) Posterior predictive p-values. The Annals of Statistics 22: 1142–1160.

[13] Consortium GP, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

[14] T1DBase. `http://www.t1dbase.org`. URL `http://www.t1dbase.org`.

[15] Nejentsev S, Howson JMM, Walker NM, Szeszko J, Field SF, et al. (2007) Localization of type 1 diabetes susceptibility to the mhc class i genes hla-b and hla-a. Nature 450: 887–892.

[16] R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

[17] Miller AJ (1984) Selection of subsets of regression variables. Journal of the Royal Statistical Society Series A (General) 147: pp. 389-425.

[18] Gring HH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69: 1357–1369.

[19] Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33: 177–182.

[20] Viallefont V, Raftery AE, Richardson S (2001) Variable selection and bayesian model averaging in case-control studies. Statistics in Medicine 20: 3215–3230.

[21] Schwarz G (1978) Estimating the dimension of a model. Ann Statist 6: 461-464.

[22] Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: A tutorial. Statistical Science 14: 382417.

[23] Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 42: 508–514.

[24] Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41: 703–707.

[25] Maier LM, Lowe CE, Cooper J, Downes K, Anderson DE, et al. (2009) Il2ra genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. PLoS Genet 5: e1000322.

[26] Martin JE, Carmona FD, Broen JCA, Simen CP, Vonk MC, et al. (2012) The autoimmune disease-associated il2ra locus is involved in the clinical manifestations of systemic sclerosis. Genes Immun 13: 191–196.

[27] Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. Nat Genet 40: 955–962.

[28] Quan C, Ren YQ, Xiang LH, Sun LD, Xu AE, et al. (2010) Genome-wide association study for vitiligo identifies susceptibility loci at 6q27 and the mhc. Nat Genet 42: 614–618.

[29] Zollner S, Pritchard JK (2007) Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80: 605–615.

[30] Sun L, Dimitromanolakis A, Faye LL, Paterson AD, Waggott D, et al. (2011) Br-squared: a practical solution to the winner's curse in genome-wide scans. Hum Genet 129: 545–552.

[31] Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. PLoS Genet 6: e1000888.

[32] Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43: 1193–1201.

[33] Hao K, Boss Y, Nickle DC, Par PD, Postma DS, et al. (2012) Lung eqtls to help reveal the molecular underpinnings of asthma. PLoS Genet 8: e1003029.

[34] Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. (2007) Genetic variants regulating ormdl3 expression contribute to the risk of childhood asthma. Nature 448: 470–473.

[35] Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, et al. (2010) A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 363: 1211–1221.

[36] Fu J, Wolfs MGM, Deelen P, Westra HJ, Fehrmann RSN, et al. (2012) Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet 8: e1002431.

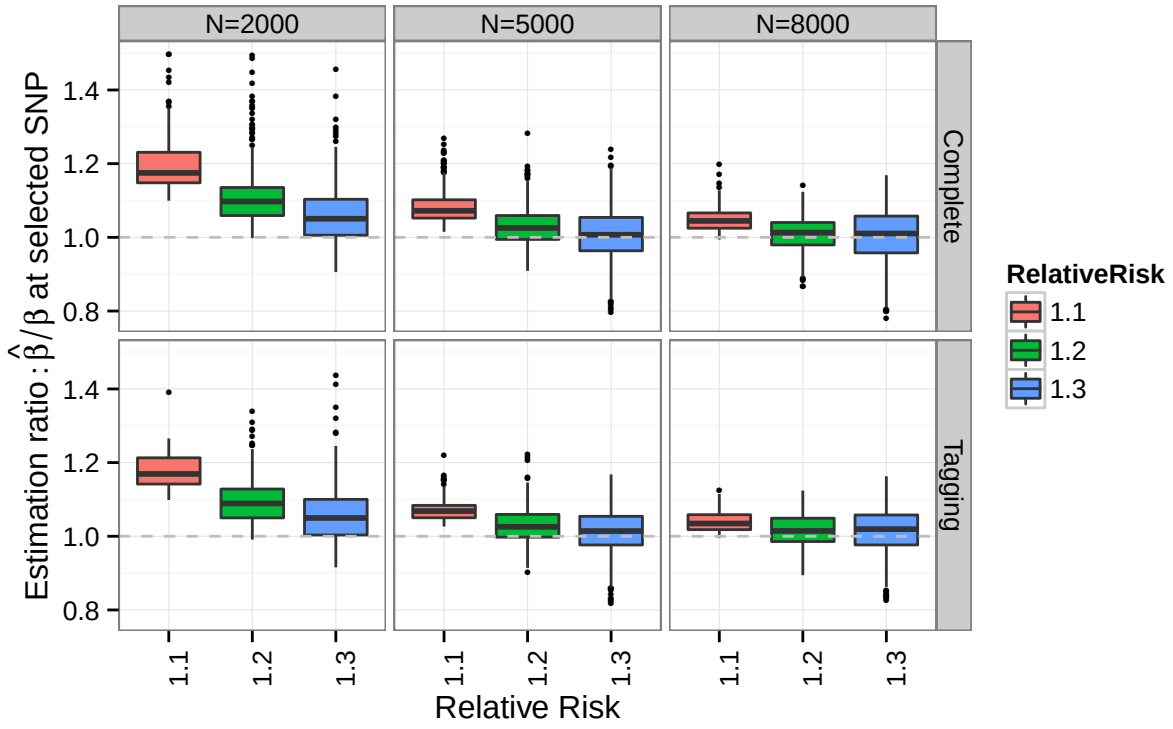


Figure 1. Effect sizes at selected SNPs tend to be overestimated. Boxplots show the distribution of the ratio of the estimated effect size ($\hat{\beta}$) to the true effect size at the selected SNP (β). Estimated effects are more likely to be biased for smaller effect sizes and sample sizes.

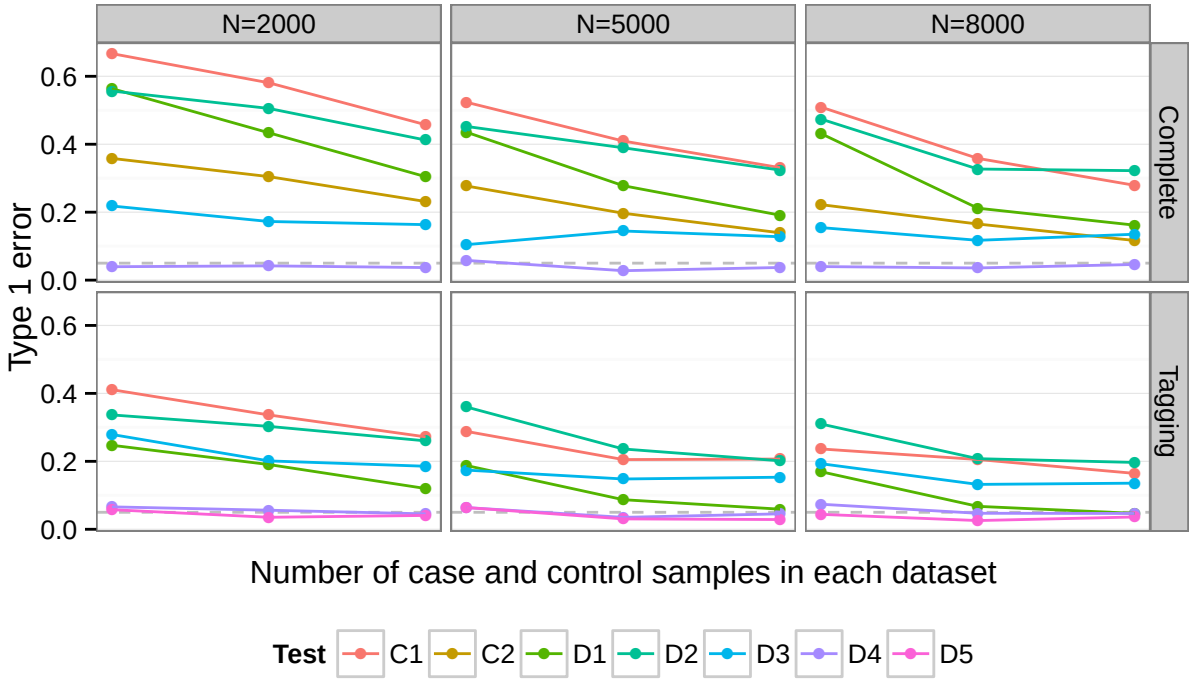


Figure 2. Type 1 error rates in naive colocalisation testing. A nominal type 1 error rate of 5% is consistently exceeded using conditional colocalisation testing conditioning on either the most associated SNP for the other trait (C1) or the common causal SNP which is only possible in simulated complete genotyping data (C2). Direct colocalisation testing tends to exhibit lower type 1 error rates, but the excess can still be substantial when using the most strongly associated SNPs in each dataset (D1); the union of lasso variable selection in each dataset (D2) or a two stage lasso variable selection (D3) as previously described [5]. In contrast , type 1 error rates are well controlled for direct testing using principle components which capture 85% of the genetic variation (D4) or within a Bayesian Model Averaging approach to variable selection (D5), even appearing conservative for small effect sizes. The X axis shows the number of cases and controls in a case-control dataset with relative risk of disease (RR) and type 1 error rates were calculated by comparing two case-control datasets of equal sample and effect size, simulated to share a common causal variant.

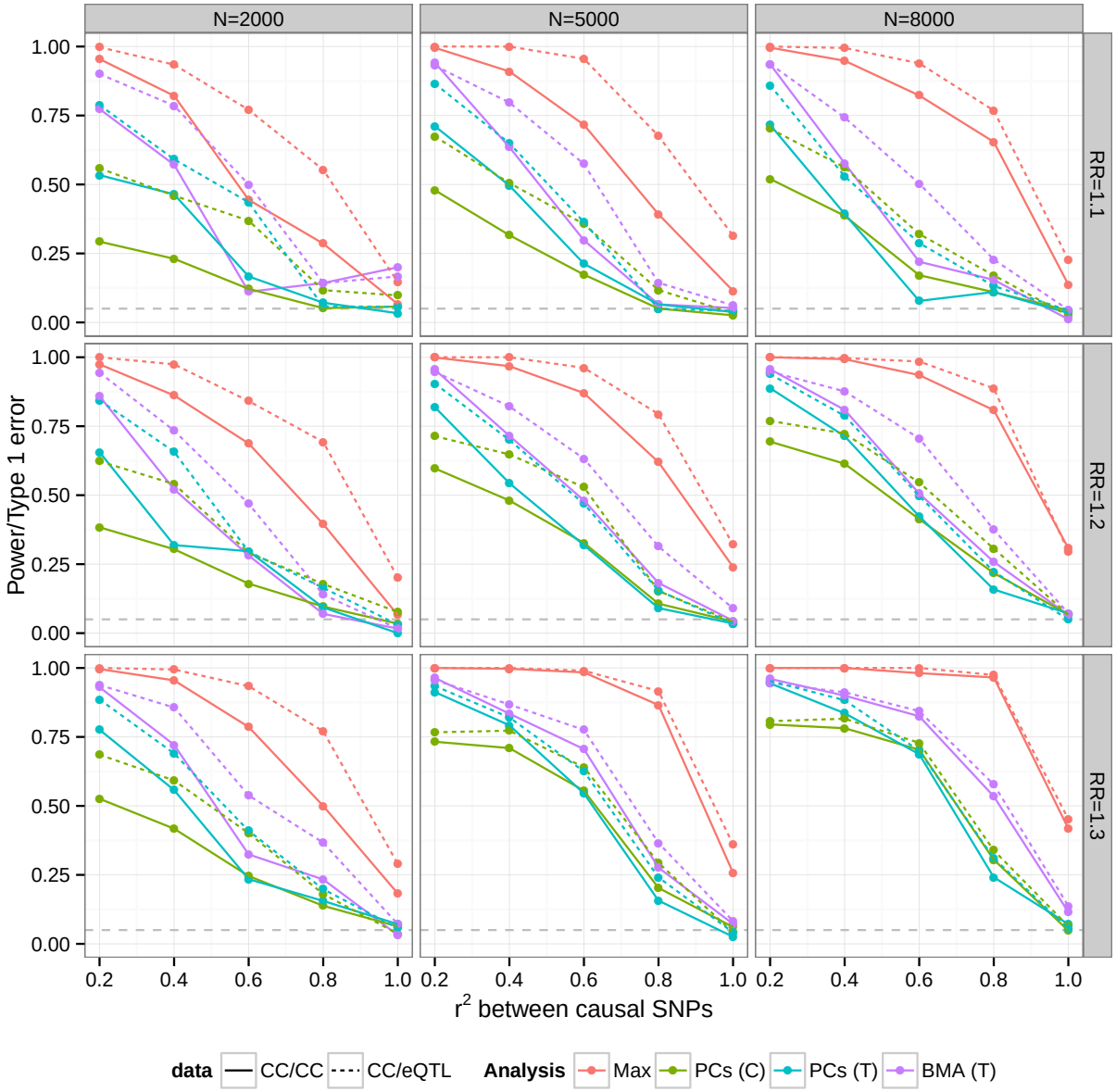


Figure 3. Power for direct colocalisation analysis using PC or BMA approaches. The theoretical maximum power (Max) is calculated by direct colocalisation testing using the two causal variants which are known in simulated data and show that the predominant determinant of power is the r^2 between the variants, with power decreasing as LD increases. When the causal variants are not known, power decreases under either a PC or BMA approach. The X axis shows the maximum r^2 between the causal variants, ie r^2 has been categorised into 5 groups: $[0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, $(0.8, 1.0]$. N is the number of cases and controls in a case-control dataset with relative risk of disease RR. Power is shown for comparing two case-control studies with equal sample numbers and effect sizes (solid lines) or for comparing a case-control study to an eQTL study of 1000 samples where the causal variant explains 30% of the variance of the expression. The PC approach was implemented by selecting the smallest subset of components which captured 85% of the genetic variance. We considered only tagging genotype scenarios to reduce computation time.

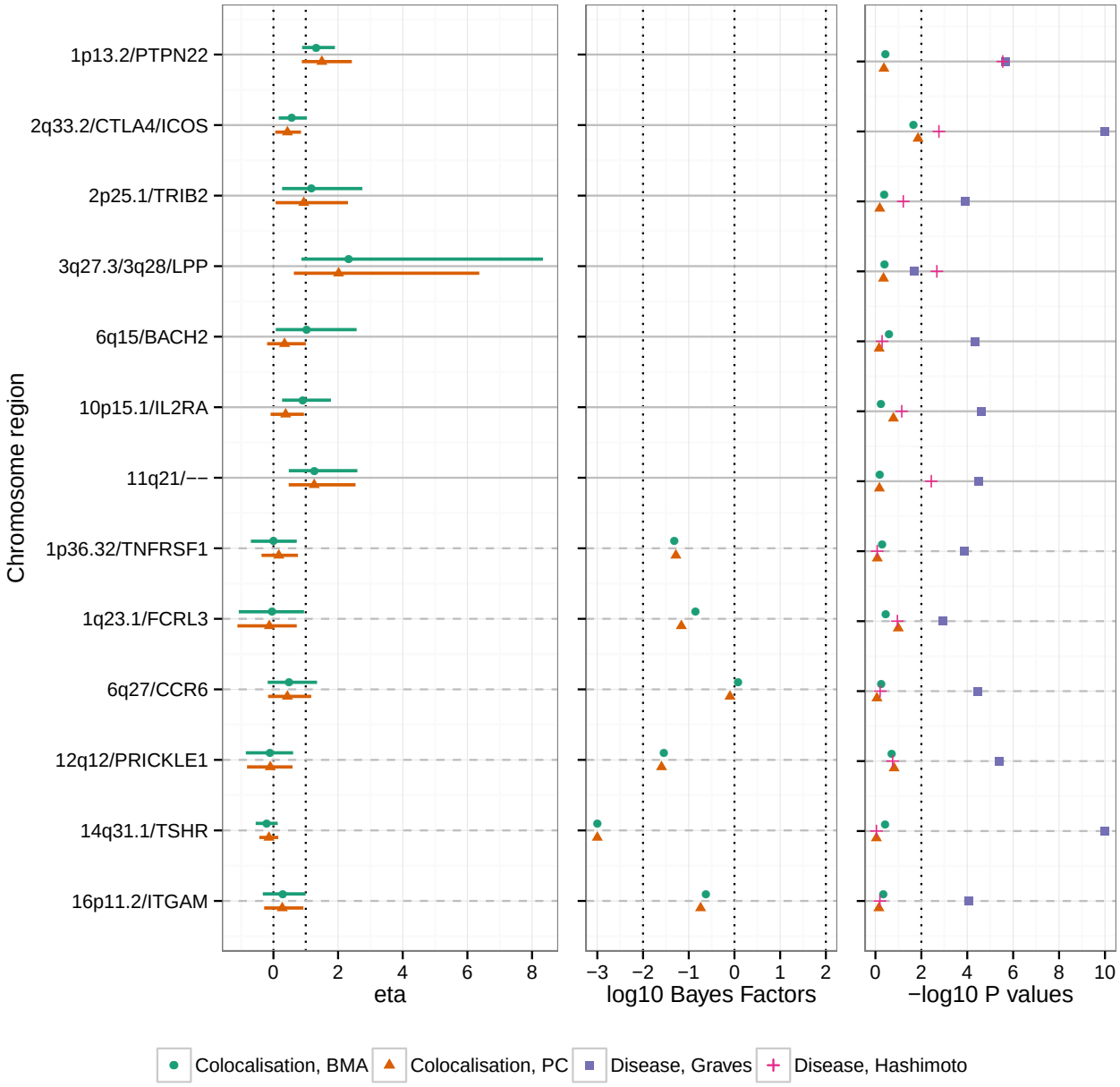


Figure 4. Colocalisation analysis of Graves' and Hashimoto's diseases. Regions are labeled by chromosome and likely candidate gene(s) and arranged so that the top seven regions showed marginally significant association with both GD and HT ($p < 0.05$) and the bottom six with just GD in the published single SNP analysis [9]. The left panel shows the estimate of the coefficient of proportionality, η, and its 95% credible interval calculated using either BMA or PC approaches. The middle panel shows the $\log_{10}$(Bayes Factor) comparing $\eta = 1$ to $\eta = 0$ for regions associated with only GD; values above 2 favour $\eta = 1$ by a ratio of 100:1 whilst values below -2 favour $\eta = 0$ by a ratio of 100:1. The right panel shows the p values for the association analysis of Graves' and Hashimoto's using the selected principal components and the posterior predictive p value for the colocalisation test. p values are shown on a $-\log_{10}$ scale, which has been truncated at 10 so that more extreme p values are displayed at $-\log_{10}(p) = 10$. For the PC approach, testing was based on the smallest subset of components that captured 90% of the genetic variance.

Supplementary Material

Simulation

Once a "causal variant" SNP, S, was selected, control haplotypes were sampled randomly and case haplotypes sampled conditional on the allele carried at S. For a disease model with relative risk r, and given the minor (risk) allele at S has frequency π_0, in controls, the frequency in cases is

$$\pi_1 = \frac{r\pi_0}{1 - \pi_0 + r\pi_0}.$$

Therefore when sampling case haplotypes, we over-sample haplotypes carrying the risk allele and under sample those carrying the protective allele by using sampling probabilities proportional to

$$P_S = \begin{cases} \dfrac{\pi_1}{\pi_0} & \text{haplotype carries risk allele} \\ \dfrac{1-\pi_1}{1-\pi_0} & \text{haplotype carries protective allele allele.} \end{cases}$$

For eQTL data, we simulated a response variable, Y as a mixture of Gaussians

$$Y = \sqrt{0.7}Z + \sqrt{0.3}X$$

where Z was sampled from a standard Gaussian and X is the count of the minor allele at the causal SNP. Thus, X would explain 30% of the variance of Y, or 30% of the simulated eQTL, independent of minor allele frequency.

The effect size at a selected SNP

To calculate the bias in figure 1, we compared the estimated effect size at the sampled SNP to the true effect *at that SNP*, ie not at the causal SNP. If the causal SNP is S and the selected SNP is T, then the underlying relative risk at T is simplest to calculate in a haploid system, which is

equivalent to assuming Hardy Weinberg equilibrium. Given

$$\rho' = \rho(S,T)\sqrt{\pi_S\pi_T(1-\pi_S)(1-\pi_T)}$$

where $\rho(S,T)$ is the correlation between S and T, then the expected proportion of cases in the population conditional on the allele carried at SNP T is

$$\begin{cases} D_1 = \dfrac{r(\pi_S\pi_T + \rho') + ((1-\pi_S)\pi_T - \rho')}{\pi_T} & T \text{ is risk allele} \\ D_0 = \dfrac{r(\pi_S(1-\pi_T) - \rho') + ((1-\pi_S)(1-\pi_T) + \rho')}{1-\pi_T} & T \text{ is protective allele} \end{cases}$$

and the relative risk is $\dfrac{D_1}{D_0}$. For a rare disease such as T1D, relative risks and odds ratios are approximately equal.

Implementation of Bayesian Model Averaging

Bayesian model averaging requires evaluating all possible multiple SNP models in each trait, and conducting colocalisation testing for each model. We began by defining the posterior probability of model j for both traits as,

$$\pi_j = \frac{\pi_j^1\pi_j^2}{\sum_k \pi_k^1\pi_k^2}$$

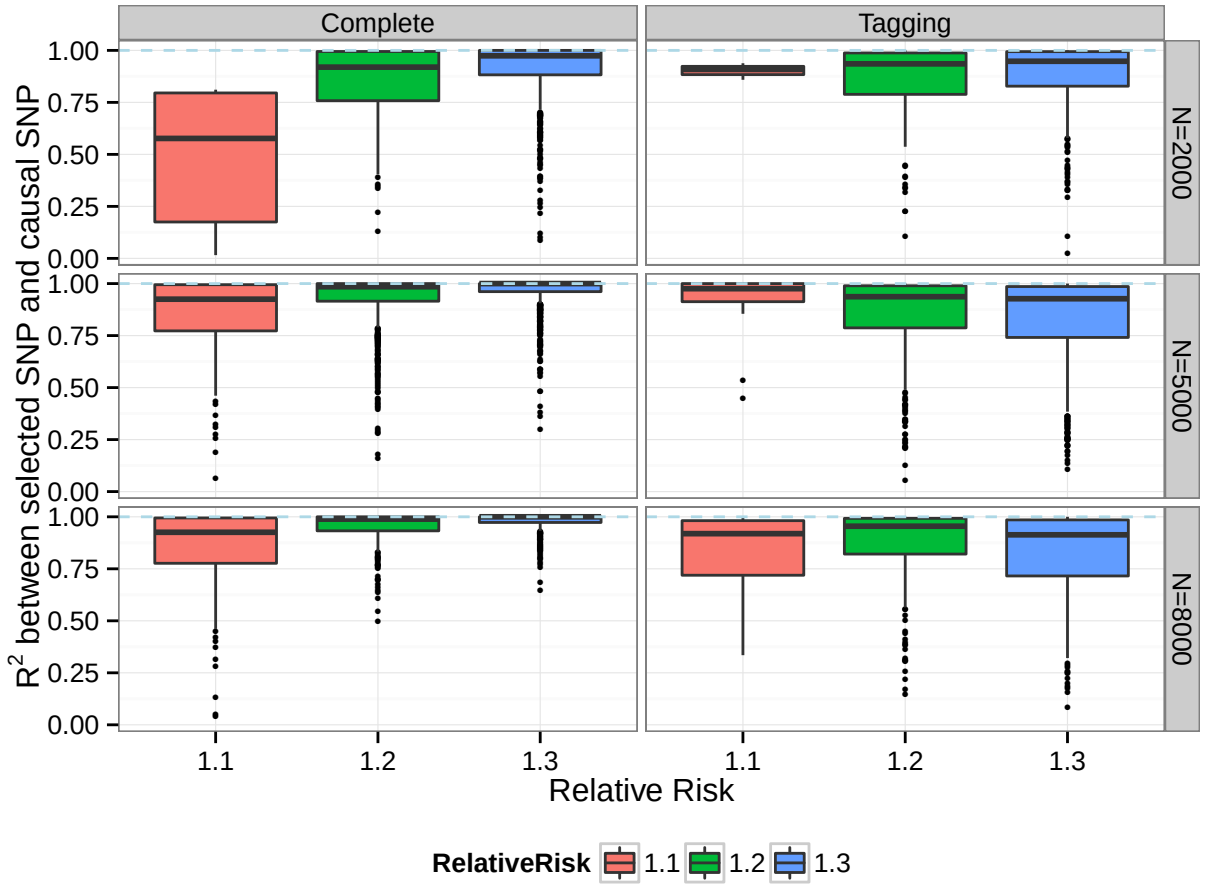
where π_j^i is the posterior probability of model j for trait i and a model j indicates which SNPs are included in the model. Even when the number of SNPs to be tested is fixed at two, the number of possible models is $\frac{p!}{2!(p-2)!}$. Whilst testing all models is feasible if computationally expensive for analysis of real data, it is impractical for simulations. To reduce the computation burden, we first evaluated all p single SNP models and indentified the set of SNPs with very low posterior probability ($\pi_j < 0.01$). We then excluded any two SNP model containing *only* SNPs from this set. If $p_0 < p$ such SNPs were identified, this reduced the number of models to test to $\frac{p!}{2!(p-2)!} - \frac{p_0!}{2!(p-2)!}$.

For the purposes of simulation, we used the profile likelihood approach to generate a χ_1^2 distributed test statistic and averaged the resulting p values, P_j, over the model space to calculate an overall posterior predictive p value, $\sum_j P_j\pi_j$. For the application to AITD, we integrated the p

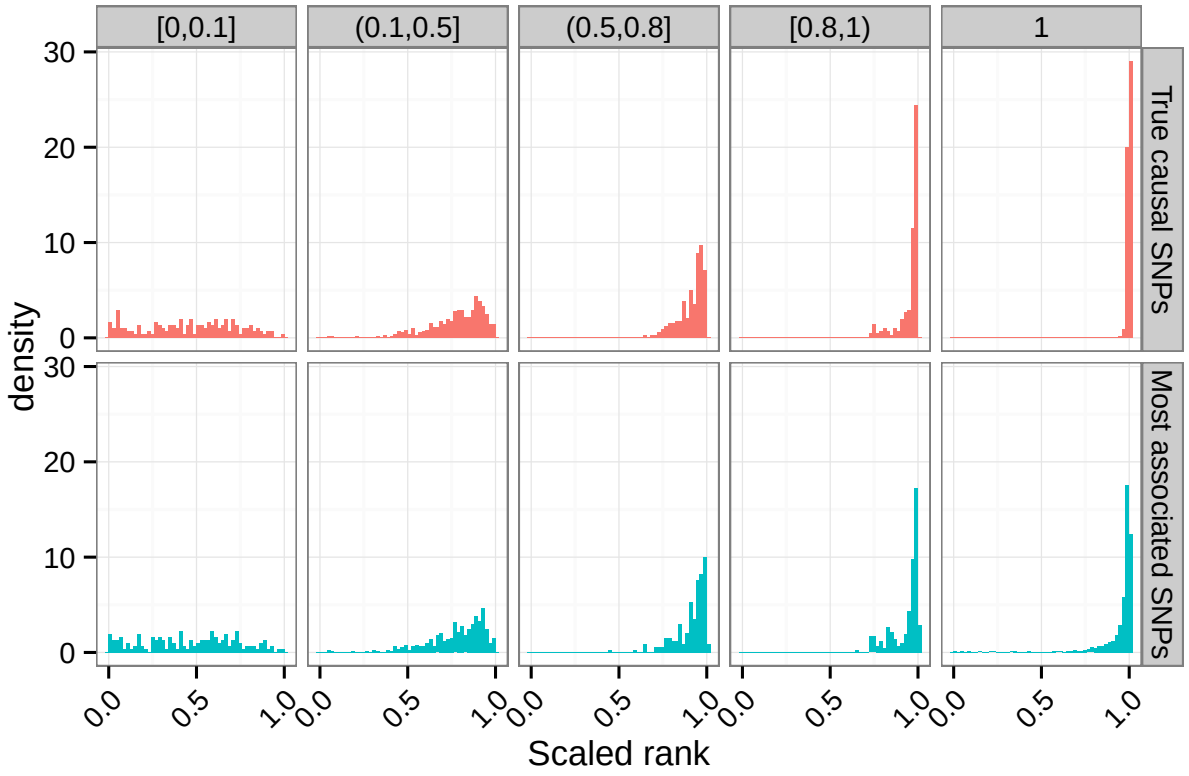
value associated with the χ^2_2 distributed test statistic calculated assuming η was known over both the posterior distribution of η given each model, and the posterior of the model space. The latter is formally correct, but comptationally too expensive for simulation, and the profile likelihood p value and the posterior predictive p value have been shown to be very similar for large samples.

Region	SNP	MAF	GD		HT		Power, $\alpha =$	
			OR	p value	OR	p value	0.05	10^{-6}
Associated with GD and HT in published study								
1p13.2/*PTPN22*	rs2476601 G>A	0.096	1.55	4.03×10^{-16}	2.02	3.74×10^{-15}	0.99	0.37
2q33.2/*CTLA4*/*ICOS*	rs11571297 G>A	0.493	0.72	2.81×10^{-23}	0.82	3.21×10^{-3}	0.99	0.490
2p25.1/*TRIB2*	rs1534422 A>G	0.455	1.16	4.69×10^{-6}	1.24	1.64×10^{-3}	0.61	0.004
3q27.3/3q28/*LPP*	rs13093110 C>T	0.452	1.18	8.17×10^{-7}	1.20	7.09×10^{-3}	0.70	0.008
6q15/*BACH2*	rs72928038 G>A	0.177	1.21	3.63×10^{-6}	1.30	1.36×10^{-3}	0.64	0.005
10p15.1/*IL2RA*	rs706779 A>G	0.467	0.85	2.27×10^{-6}	0.84	0.0125	0.674	0.007
11q21/–	rs4409785 T>C	0.173	1.21	5.37×10^{-6}	1.34	3.54×10^{-4}	0.63	0.004
Associated with GD only in published study								
1p36.32/*TNFRSF1*	rs2843403 C>T	0.362	0.84	7.94×10^{-7}	0.97	0.696	0.69	0.007
1q23.1/*FCRL3*	rs7522061 T>C	0.480	1.16	1.08×10^{-5}	1.03	0.634	0.60	0.004
6q27/*CCR6*	imm_6_167338101 A>C	0.408	0.84	3.30×10^{-7}	0.88	0.056	0.71	0.009
12q12/*PRICKLE1*	rs4768412 C>T	0.363	1.19	3.30×10^{-7}	1.00	0.949	0.73	0.010
14q31.1/*TSHR*	rs2300519 T>A	0.380	1.54	1.34×10^{-38}	0.93	0.295	1	0.95
16p11.2/*ITGAM*	rs57348955 G>A	0.396	0.83	3.76×10^{-8}	0.91	0.188	0.76	0.013

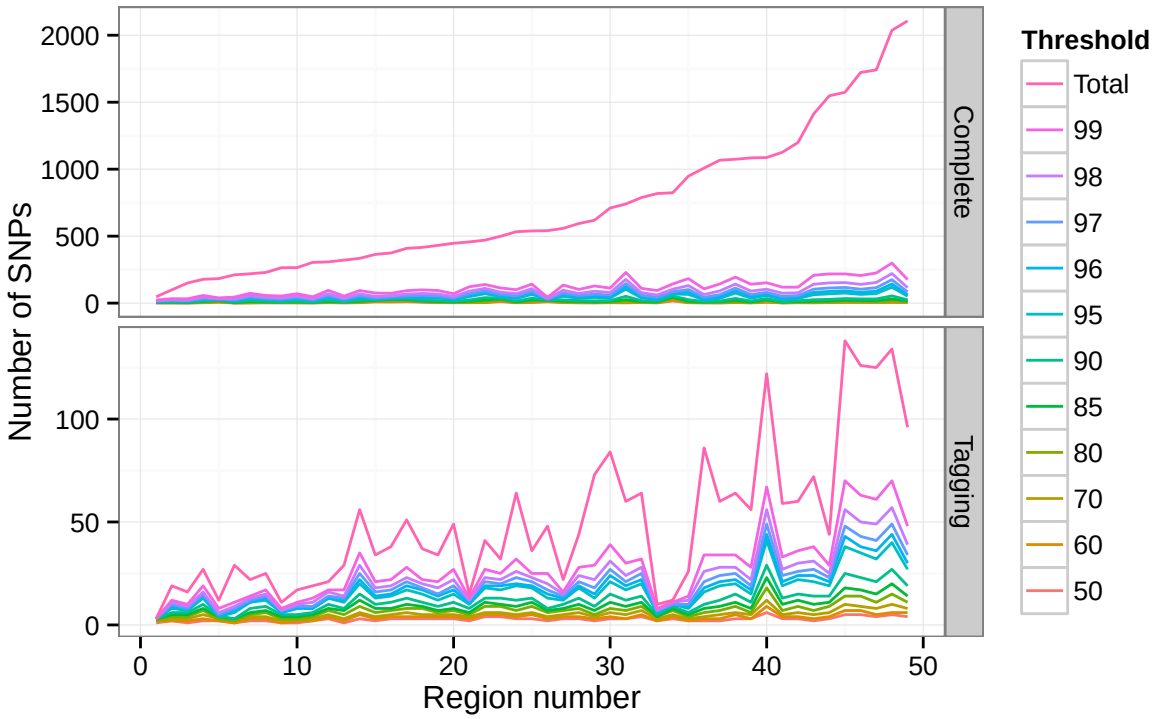
Supplementary Table 1. Power to detect association with Hashimoto's Thyroiditis to confirmed loci for Graves' Disease. Region denotes chromosomal region and most likely candidate gene(s) where available [9]. GD=Graves' Disease; HT=Hashimoto's Thyroiditis; MAF=minor allele frequency in controls.



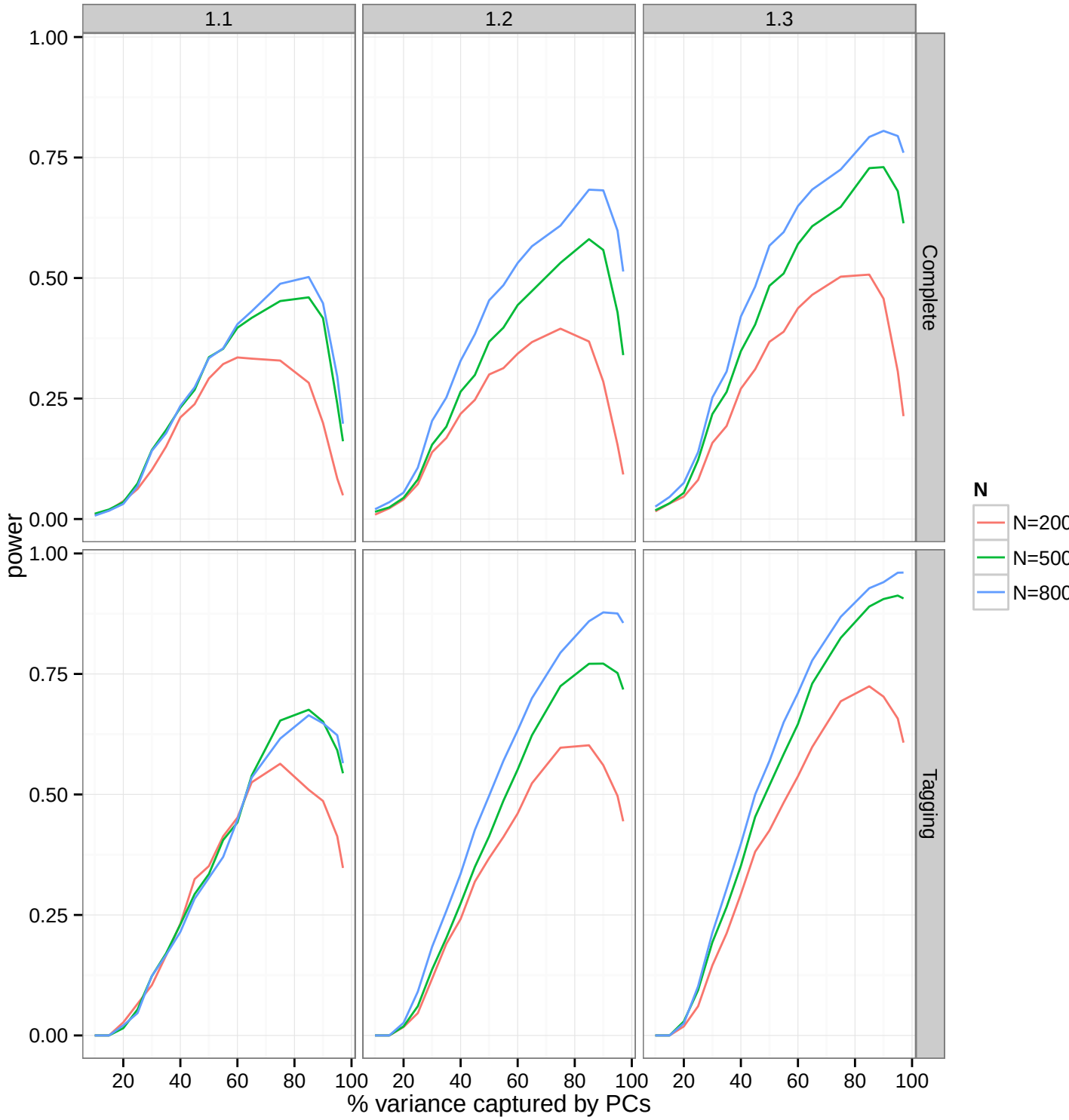
Supplementary Figure 5. The most associated SNP in a region is not necessarily the causal SNP. Boxplots show the distribution of r^2 between the SNP with the smallest p value (conditional on $p < 1 \times 10^{-8}$) and the causal SNP from simulated data, either under tagging or complete genotype coverage. Increasing the effect size increases the range of tagging SNPs detectable, and hence can have the apparently counter-intuitive result of decreasing the correlation between selected and causal SNPs. However, if complete genotype coverage is available, the LD between selected and causal SNPs tends to increase with effect size or sample size.



Supplementary Figure 6. Distribution of Nica *et. al*'s rank statistic. The statistic is evenly distributed within [0,1] when the LD between the causal variants is negligable, but is increasingly biased towards 1 as the LD increases. Columns are divided by the r^2 between distinct causal variants, with $r^2 = 1$ indicating a shared causal variant. The top row shows the optimal result that could be obtained if conditioning on the true causal variant were possible, the bottom row shows the effect of conditioning on the most associated SNP is to reduce the degree of skew. Results are shown for a complete genotyping scenario, with a sample size of 2000 and a relative risk of 1.3. Similar effects are seen under tagging or complete genotyping approaches, but the skew towards 1 occurs more rapidly with larger samples and effect sizes.



Supplementary Figure 7. The number of principal components required to capture a predefined proportion of the variance. The 49 regions used for simulation are displayed, unlabelled and ordered by the total number of SNPs. The majority of variation can be captured by a relatively modest number of components even for regions containing large numbers of SNPs. Threshold specifies the minimum proportion of variance captured, or "Total" for the total number of SNPs in a region.



Supplementary Figure 8. Power using colocalisation testing of principal components according to the proportion of genotype variance captured. Power is shown for all simulated datasets where the r^2 between the causal SNPs was less than 0.5.